nature medicine

Review article

https://doi.org/10.1038/s41591-026-04351-4

Global approaches to infectious disease surveillance and modeling

Received: 11 July 2025

Accepted: 17 March 2026

Published online: 20 May 2026

Check for updates

A list of authors and their affiliations appears at the end of the paper

Human mobility, climate change and demographic trends increase the risk of pathogen spillover and expansion. Data that can inform our responses to outbreaks have increased in availability and volume, but access to highly confidential outbreak data and commercially sensitive contextual information remains difficult. Despite ongoing efforts to adopt global health data infrastructures and sharing protocols, there remain regulatory, logistical, human and computational barriers to data sharing. Federated approaches—in which data remain stored locally but analyses are performed across datasets from different sources—offer a potential way to address these challenges. While federated approaches have been used in some clinical and biomedical contexts, their adoption in infectious disease surveillance and modeling has been limited. Here, we discuss global approaches to infectious disease modeling and analysis, with a focus on federated methods. We outline how these can be used to address key epidemiological questions during outbreaks by enabling the secure use of multimodal data and integration with existing surveillance and modeling efforts. We summarize current methods for combining distributed and locally stored data and identify limitations, opportunities and organizational structures needed to achieve equitable global public health impacts.

The increasing speed at which high-quality data can be collected during infectious disease outbreaks provides new opportunities for understanding and analyzing epidemics[1]. Data generated during an outbreak might include patient-level data (for example, line lists, clinical records with recorded symptoms, exposure history, serology, virology, vaccinations, treatments and preexisting conditions), pathogen and human genomic data, and contextual data (for example, information relating to mobility, contacts, socio-economics, interventions, behaviors, and livestock and wildlife health). However, these data are often stored in disparate noninteroperable formats, collected by different agencies and private companies, and may have complicated access permissions and/or physical storage requirements[2–4].

Coanalyzing increasingly disparate data sources across locations can yield insights into the origins and transmission dynamics of outbreaks, as well as the effectiveness of interventions. While some combinations of the abovementioned data types have been used to investigate single-country or regional outbreaks[5–8], no system that enables broader adoption currently exists[2,9]. Difficulties in timely data access and interoperability continue to impede analyses, affecting the timeliness and effectiveness of responses to outbreaks of COVID-19, mpox, Ebola, Marburg and highly pathogenic avian influenza H5N1, as well as to seasonally recurring infections[2,10–12]. Continued monitoring of pathogen evolution could accelerate the detection of emerging variants and their geographical spread.

During disease outbreaks and epidemics, sharing data is sensitive due to concerns about reidentification of infected individuals, stigma, rules surrounding ownership and other factors (for example, hospital networks often own patient data). Political and economic considerations can also influence the willingness of parties to share data, as reporting outbreaks may affect international travel, trade or a country's perceived ability to control an epidemic.

Addressing these capacity, regulatory, ownership, sovereignty and privacy challenges necessitates new analytical approaches. In this context, federated approaches have emerged as a promising paradigm.

e-mail: s.scarpino@northeastern.edu; samir.bhatt@sund.ku.dk; moritz.kraemer@biology.ox.ac.uk

Table 1 | Key components of a federated system for epidemic surveillance and modeling

Component	Function	Examples
Local data nodes	To store locally generated, confidential datasets	At the individual level, behavioral data from mobile phones[3] or clinical data from hospitals[14,17]
Global data nodes	To store and share globally available data	International flight data (commercial)[26] or satellite climate data[96,97]
Federated model or algorithm	To extract meaningful summary statistics from locally stored data	Estimation of variant secondary attack rates at the household level[157]
Local computation	To execute analyses at the local level in secure server environments, potentially using information from global nodes	Local analysis of epidemic data
Aggregation layer	To summarize and aggregate local analyses at a central node	Combining results from multiple studies using evidence synthesis or model averaging[40]
Privacy-preserving technology	To ensure that the system is compliant with local and international laws and ethical standards	Differential privacy[123,130] or digital twinning[128]

We define federated approaches broadly as any system in which analyses are performed across datasets from different sources, but the raw data remain stored locally and are not exchanged centrally[13] (Table 1). A special use case is federated learning, in which a machine learning model is trained under the orchestration of a central server[13]. Unlike traditional collaborative analyses, federated architectures allow sites to perform dependent and iterative computations across datasets, integrating heterogeneous data in near real time while maintaining privacy and data sovereignty. These approaches have been adopted elsewhere in healthcare[14–18] (where datasets from individual hospitals are small) and in mobile and edge devices (where data packages could be exposed to attacks during transfer)[14,19,20].

The use of federated models in infectious disease and outbreak research has been limited to date, but technical developments have created new opportunities and potential applications. There are many potential uses of federated data structures in the field of infectious diseases, including the following: monitoring the emergence of new pathogens[21] through sentinel surveillance without sharing raw data from each site; studying transmission dynamics and key epidemiological parameters in highly connected countries to understand how outbreaks are related; evaluating vaccine effectiveness across subpopulations; predicting clinical outcomes across health systems; and assessing the impact of countermeasures[22,23].

In this Review, we discuss how global approaches to infectious disease analysis and modeling, particularly federated methods, could reduce the time needed to generate actionable inferences during infectious disease outbreaks[21]. We describe federated architectures to inform real-time deployment across diverse data types (for example, point-of-care molecular diagnostics, genomics, metagenomics, clinical data and health systems surveillance) and outline an approach for integrating multimodal data and artificial intelligence (AI) using privacy-preserving infrastructure, addressing long-standing challenges in data interoperability, equity and trust.

Global data sharing

The COVID-19 pandemic response saw an unprecedented expansion of global capacity for pathogen data generation, analysis and international collaboration, with investments in sequencing infrastructure, analytics and data platforms enabling global situational awareness in near real time[24]. At the same time, the pandemic underscored that sustainable national and regional surveillance capacities depend on maintaining data and analytical sovereignty, including the rights, recognition and agency of data generators and stewards[25]. Pathogen data have long been shared through multiple modalities, ranging from open repositories to registered-access, controlled and locally governed systems. This reflects the varying sensitivities of different data types, legal and ethical obligations, and the incentives and risk perceptions of data holders[25].

Rather than converging on a single uniform model, current efforts increasingly focus on strengthening a global data-sharing ecosystem that accommodates this diversity while improving interoperability, attribution and transparency across local, national and global levels of surveillance. Analyses of the COVID-19 pandemic demonstrate how federated approaches could have enabled real-time estimation of key epidemiological parameters and genomic surveillance across institutions and regions without requiring the centralization of sensitive data[5,6,26,27].

This approach has been further articulated by the World Health Organization (WHO) through its guiding principles for pathogen genome data sharing[28] and its attributes and principles for pathogen genomic data-sharing platforms[29], which define the technical, operational and ethical foundations for transparent, accountable and equitable data exchange in support of public health. Within this framework, federated architectures could represent an important component of modern data sharing, enabling decentralized analysis and collaboration while respecting legal, ethical and institutional constraints on data movement.

Applications of federated approaches in epidemic modeling

Accurate inference of key epidemiological parameters during the early stages of outbreaks is critical for informing effective policy decisions. They can help determine isolation and follow-up periods after known exposures or the onset of symptoms. Parameters of particular relevance include the incubation period (the time from exposure to symptom onset), serial interval (the time between the symptom onsets of two successive cases), infectious period and fatality rate; other inferences can identify potential geographical spread and populations at highest risk. However, the limited number of early cases often prevents robust inference of these parameters. Early case counts are typically biased toward more severe infections, as asymptomatic or mildly symptomatic cases are often missed, which can delay the recognition of key transmission dynamics[30,31]. During the 2022 global mpox outbreak, country-level data were too infrequent to accurately estimate incubation periods[32]. During the COVID-19 pandemic, when new variants emerged, early inferences of their incubation periods were often imprecise.

Further, estimating the relative growth rates of emerging variants of concern, their clinical severity and their mechanisms of spread requires the integration of multiple data sources (genomic, epidemiological, immunological, experimental and laboratory data, as well as contextual information) that are collected and owned by different agencies, including the private sector[27,33–35]. It is not only that the data are necessarily limited, but also that they are distributed across multiple sites and lack standardization.

Federated approaches address these challenges in some contexts, but their application varies by pathogen and setting. They are likely to be useful when data are abundant and case trajectories can be easily compared across regions. In data-poor situations, federated approaches are particularly useful because parameters can then be informed by data from multiple settings. For vector-borne diseases (for example, dengue and malaria), federated approaches can help identify shared climatic drivers across regions. However, for pathogens primarily driven by environmental conditions (for example, leptospirosis), the benefits of federated learning may be more limited, as transmission

https://doi.org/10.1038/s41591-026-04351-4

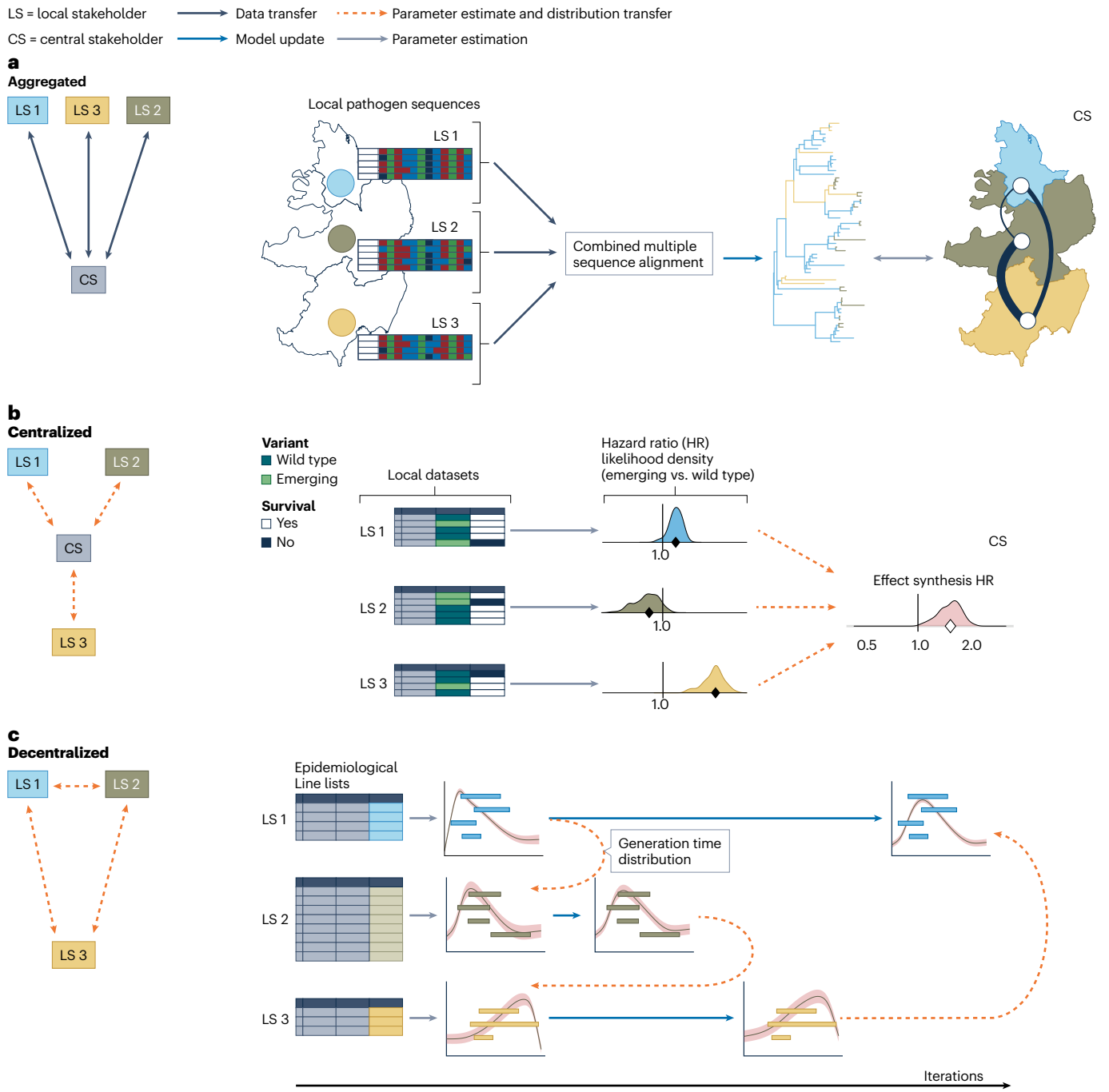

Fig. 1 | Epidemiological analyses and the degree of decentralization. a, An aggregated structure illustrated by a phylogeographical analysis quantifying the number of pathogen movements between three geographical locations. Each local stakeholder produces its own local data, which are shared with a central stakeholder that combines the genomic sequence data and carries out the analysis. **b**, A centralized structure illustrated by a pipeline for estimating the relative hazard ratio for death caused by an emerging variant, relative to a wild-type variant. Local stakeholders generate local estimates with their unique likelihood densities, which are shared with a central stakeholder that produces a global estimate through effect synthesis meta-analysis. **c**, A decentralized structure illustrated by the estimation of a pathogen's incubation period. Local stakeholders fit models to locally collected data, and fitted model parameters are shared sequentially between stakeholders to update the model fits and parameter estimates. The workflow can accommodate the inclusion of new data in local stakeholder datasets.

dynamics depend less on cross-border data integration and more on local ecological surveillance. A variety of federated architectures exist that can accommodate the complex and varied epidemiological data and analysis requirements, as discussed below.

Architectures for epidemic modeling

Existing federated frameworks could be adapted to fit the requirements of infectious disease modeling (Fig. 1). Phylogenetic analysis to track the origins or relatedness of outbreaks across regions will require multiple sequence alignments to be assembled in a centralized manner for analysis, before results are returned to the networks of contributors (Fig. 1a). Widely used phylogenetic data repositories have access restrictions that limit real-time analysis, meaning insights are often generated retrospectively rather than contemporaneously. In such cases, a comanaged platform could be used to perform standardized data analysis within trusted research environments with transparent

governance, supported by clear principles for access, accountability, interoperability, timeliness and equitable benefit sharing, which are increasingly recognized as essential for genomic surveillance systems[29]. Cross-border surveillance efforts would particularly benefit from this model, as geographical proximity and shared disease risks—such as those posed by viral hemorrhagic fevers—create strong incentives for collaboration and data sharing between highly connected countries (Box 1). Current initiatives addressing data sharing during outbreaks include Pathoplexus[36] for decentralized genomic data storage and Global.health[37] for anonymized epidemiological data sharing. For federated approaches specifically, the European Genomic Data Infrastructure project recently demonstrated a federated infrastructure that allows a user in one country to analyze synthetic genomic data across other national nodes without the data leaving the country of origin[38]—a model that aligns with emerging data governance frameworks such as the European Health Data Space, which mandates nationally hosted data with controlled access points[39].

To estimate the clinical severity and risk factors for an emerging infectious disease during its early phase, we can imagine that local analyses of highly confidential data are performed either at the hospital site or within the local healthcare database, before their inferred parameters are combined using evidence synthesis or model averaging by a central stakeholder[40–42] (Fig. 1b). We take inspiration from the Observational Health Data Science and Informatics (OHDSI) collaborative[43], in which data partners manage more than 450 independent administrative claims and electronic health record databases, collectively covering nearly 8% of the world's population. Through open-source analytical tools[44] based on a common data model[45], partners participate in coordinated global observational studies that improve collaborative analysis while limiting direct data sharing[46,47]. To date, OHDSI does not accommodate central coordination, which would enable more extensive and iterative methods in federated learning, especially for datasets collected using different protocols[20].

In the future, more decentralized models could take individual data from each site (Fig. 1c), fit distributions or parameters, and then share their posterior distributions with the other sites[23,48]. Each site can then update its estimates using these distributions as inputs for the local model. As with other approaches, this requires coordination between stakeholders but can be performed asynchronously.

Advanced analytics for federated modeling

Modeling disease dynamics across geographical locations can be achieved using a range of advanced statistical methods. Federated learning methods[13,49,50] could be used to analyze multicenter, hierarchical or longitudinal data (such as patient outcomes), with iterative improvements as new data are collected. Transfer or self-supervised learning could leverage pretrained models from similar domains or past outbreaks for new downstream tasks, such as predicting the geographical spread of pathogens and combining models (embeddings) with locally collected mobility and behavioral data[51–57].

However, a central challenge in federated learning is the often non-independent, nonidentically distributed nature of decentralized data[58], which may be biased or partially correlated (for example, disease prevalence or demographic differences across space and time). This can lead to biased estimates and poor generalization in diverse real-world settings, as models trained on data from one region or group may not perform well on other data with different characteristics[13,59–61]. Adaptive optimization techniques[62–64] for regularizing model updates and personalized federated learning[65–67] for local model customization (while maintaining a shared global model) have been proposed as promising future avenues. Another common approach is to reweight data points depending on the target region, with these weights shared alongside model updates for continuous improvement[68,69] (Table 2). For contextual data (such as human mobility information), a technique known as post-stratification assigns weights to data based on the demographic profile of a given location (for example, census tract age–sex strata, urbanicity and income quintile). Classic propensity-score techniques align the study population with population totals and have been shown to reduce bias and improve external validity in public health surveys and cohort studies[70,71].

Beyond federated learning, other approaches combine mechanistic modeling with machine learning to support inference from

BOX 1

Tracking the cross-border spread of emerging infectious diseases

Identifying the original location and timing of new outbreaks is critical for determining the current and future risks of the disease. A widely adopted framework now aims to detect an outbreak within 7 days, notify authorities within 1 day of detection and initiate a response within 7 days of notification[21]. However, the timeliness and accuracy of inferences about outbreak origins and spread depend on the quality and representativeness of the input data, with spatiotemporal biases in genomic sampling potentially distorting conclusions about pathogen spread[27,33,34]. In addition, different institutions often control separate data segments during outbreaks; sharing these data across institutions and borders is complex and requires cumbersome legal agreements.

These challenges underscore the importance of robust frameworks for managing and integrating diverse datasets. Recent large-scale initiatives, including CLIMB-COVID[201], Terra[202], Pathogenwatch (https://pathogen.watch/) and Nextstrain[203], have aimed to standardize genomic analysis ingestion and workflows. For example, CLIMB-COVID uses a decentralized approach for sequence data ingestion but keeps computational analyses centralized—a model further adopted by the mSCAPE infectious disease metagenomics surveillance network.

For phylogeographical analyses, a potential model for assessing pathogen spread could involve building a compartmentalized sharing model of intermediate outputs. Here, local genomic data are used to build a preliminary phylogenetic tree by incorporating private and publicly available genomes to produce an anonymized mutation-annotated tree (MAT)[204] (Fig. 2). This MAT is then shared with researchers who have full access to the source location's genomic database. In turn, these researchers perform sequence insertion (for example, using UShER[204] or MAPLE[205]) to refine the phylogenetic tree and identify independent viral introduction events that might have been obscured[204–206]. Independently, air traffic and incidence data can be used to estimate importation intensity, which can then be compared to importations inferred from genomic data alone.

Another model could involve maintaining a centralized MAT, where users locally reduce their sequence data to mutation information and submit this to a central server without including any sensitive metadata. In this framework, users receive the placement of their genome within the tree, and the centralized MAT can be updated incrementally, with sharing controlled according to user preferences. Once identified, each subtree corresponding to an introduction event can be extracted and shared with local researchers for further analysis. This could be extended to include individual travel history data, thereby improving the identification of local transmission lineages.

https://doi.org/10.1038/s41591-026-04351-4

Table 2 | Statistical methods and frameworks for infectious disease modeling

Problem domain	Statistical method	Application domain	Potential impact and obstacles
Parameter estimation from limited data	Bayesian model averaging	Inference of epidemiological parameters (for example, incubation periods)	Addresses issues of limited data per location and inconsistencies in methods applied across settings; improves the comparability of parameter estimates. Challenges include limitations with complex models (for example, latent variables) and the need for standardization across sites to prevent errors in localized models[40].
	Meta-analysis	Inference of epidemiological parameters (for example, case fatality rates)	Synthesizes results from multiple studies, field epidemiology and contact tracing to improve precision and provide more reliable estimates by pooling data. Challenges include heterogeneity in study designs and difficulty in accounting for variability between studies[158–160].
	Distributed stochastic process models (for example, Gaussian processes)	Modeling disease transmission dynamics with uncertainty estimation on small datasets	Infer robust relationships between disease dynamics (for example, incidence, prevalence and risk factors). Challenges include data scarcity, parallel computation and computational complexity[41,42].
Epidemiological modeling	Probabilistic models (for example, generative Bayesian models)	Dynamic forecasting and spatiotemporal modeling	Improve accuracy, real-time predictions and outbreak tracing; address spatiotemporal challenges arising from changes in administrative boundaries. Challenges include scaling, data integration (for example, compatibility) and computation[161–164].
	Approximate inference (for example, variational inference)	Scalable Bayesian inference for phylogenetic and mechanistic infectious disease models	Efficient phylogenetic inference and scalable analysis of large datasets, including outbreak detection and variant tracking. Challenges include balancing approximation accuracy and computational efficiency[165–171].
	Distributed deep learning (for example, transfer learning and self-supervised learning)	Epidemic nowcasting and forecasting using small datasets	Develops pretrained models for predicting epidemiological parameters (for example, case counts); summarizes complex multimodal data into linear features; estimates R_0 values from phylogenies; and detects diseases from electronic health records for related but new pathogens. Challenges include determining whether pathogens are similar enough for effective transfer learning, addressing temporal shifts and spatial heterogeneities in data collection, and mitigating reduced accuracy in later outbreak stages[51–57,172].
	Bayesian optimization	Enhancing prediction accuracy in complex models, particularly in resource-constrained environments	Accelerates parameter learning and hyperparameter optimization while improving epidemiological modeling (for example, SIR (susceptible, infectious and/or recovered) models and disease forecasting) and calibrating complex simulations (for example, malaria and SARS-CoV-2). Challenges include computational and memory demands, difficulty handling high-dimensional information and exploration–exploitation trade-offs[173–177].
	Differentiable and private agent-based models	Modeling individual behaviors, policy evaluation and proactive contact tracing (for example, personalized risk estimation)	Enable rapid, population-scale simulation and calibration with privacy-preserving decentralized computation; enhance contact tracing apps for proactive outbreak management. Challenges include modeling complex individual decisions and ensuring representative simulations[72–74].
	Bayesian spatial models	Mapping disease risk and environmental drivers (for example, climate and vectors)	Quantify the relationship between disease risk and environmental drivers (for example, climate and land use) by learning model weights across distributed datasets without exchanging raw location data. This could allow sites with sparse surveillance to improve their local risk mapping by using parameters learned from data-rich regions. Challenges include the high computational cost of spatial analyses and harmonizing disparate spatial resolutions[178,179].
Multilevel and causal learning	Federated linear mixed and hierarchical modeling	Analysis of multicenter, hierarchical or longitudinal data (for example, patient outcomes or viral evolution), inferring risk factors and outcomes	Enables robust evaluation of interventions (for example, vaccines) and risk factors (for example, comorbidities) for clinical outcomes (for example, mortality), while preserving local trend variation across sites. Challenges include risks of data tampering and hidden biases, balancing local trends with overall accuracy, and managing the high computing demands of secure, large-scale collaboration[180–185].
	Distributed causal inference	Estimation of causal effects in infectious disease modeling and public health interventions	Enables robust evaluation of interventions (for example, vaccines) and risk factors (for example, comorbidities) for clinical outcomes (for example, mortality). Challenges include handling data heterogeneity, missing data, the computational complexity of scaling causal models (for example, confounder adjustment and uncertainty estimation) to large multicountry datasets, and the lack of counterfactuals due to the observational nature of the data[186–190].
	Neural networks	Broad range of applications ranging from epidemic forecasting to clinical treatment algorithms	Evaluate the effects of policies (for example, nonpharmaceutical interventions); develop clinical treatment algorithms; enhance epidemic forecasting and geospatial modeling; and improve generalization and robustness (for example, ensemble methods including stacking). Challenges include risks of data tampering and hidden biases, ensuring fairness across different sites, balancing local trends with overall accuracy, and managing the high computing demands of secure, large-scale collaboration[184,185,191–197].
	Distributed deep reinforcement learning	Accelerated deep learning training across distributed networks	Improves clinical decision-making, survey design, real-time disease detection and intervention strategies. Challenges include high computational demands and balancing exploration–exploitation trade-offs[198–200].

complex, decentralized datasets. One such approach is the use of differentiable agent-based models (Table 2), which integrate traditional epidemiological simulations with neural networks to model individual-level behavior during outbreaks[72–74]. These models enable efficient gradient-based calibration without the need for computationally expensive surrogate models, allowing them to ingest heterogeneous data (for example, mobility patterns and wastewater surveillance[75–78]) for rapid calibration[73,79]. Recent advances support calibration across multiple institutions by splitting the modeling architecture: neural networks at local sites process sensitive data into embeddings, which are shared with a central server hosting the agent-based model[80].

At the individual level, digital contact tracing offers a decentralized way to collect the necessary data to estimate key disease parameters, such as incubation and latent periods, and provides insights for individual-level and public health decisions[81,82]. Apps can collect mobility and contact data directly on users' phones and can perform decentralized inference using message-passing algorithms[83]. However, current implementations do not support federated analytics such as exposure management, intervention evaluation or risk minimization. Private agent-based models can address this challenge by running simulations and calibration directly on personal devices, where individual contact tracing information resides[84]. This approach allows contact tracing apps to provide proactive, personalized guidance (for example, "How can I reduce my exposure risk?") without compromising data privacy. While many of these methods[73,79,85] are increasingly used in statistics and machine learning, their application to infectious disease surveillance and modeling remains nascent. These models are also likely to be most informative when calibrated with empirical contact tracing data, reinforcing the critical role of field epidemiological investigations in understanding early outbreak dynamics.

While fully decentralized models without any data sharing might work in some contexts, we cannot overstate the importance of sharing pathogen genome data for the global tracking of infectious diseases. The rapid generation and sharing of genomic sequence data during the COVID-19 pandemic enabled researchers to track the emergence of new variants of concern by comparing their datasets to those collected elsewhere[86]. This was made possible only because central stakeholders maintained large phylogenetic trees that enabled local users to determine whether their locally generated genomes fell into predesignated lineages and to evaluate their potential public health impact. This general framework has now been adopted for other pathogens (for example, monkeypox or dengue virus)[87,88] and will continue to require timely data sharing. However, a peer-to-peer system could be implemented to identify key mutations of importance, which would only require an understanding of sequence similarity. Integrating embeddings from pretrained AI models into the assessment of the pandemic potential of newly identified pathogens (using metagenomics, for example) could enable rapid risk assessments with limited need for data sharing[89] (see details in Box 2 and Fig. 2).

Lastly, it is essential to assess whether federated learning and decentralized data analysis offer epistemic advantages by balancing potential knowledge gains against the risks of privacy breaches, while ensuring that the benefits outweigh the resource costs involved, such as those required to run the analyses. Innovations in blockchain-based federated learning have the potential to overcome a number of data security concerns by ensuring the verifiability and auditability of models[90]. Approaches that consider the added value of federated analyses, such as those based on information theory, can be used to help determine the benefits of these analyses and, ultimately, their investment value[91,92]. Developing benchmark datasets tailored to unique data types (for example, genomic or metagenomic data) may be necessary to further clarify when federated models are most impactful.

BOX 2

Predicting the pandemic potential of new pathogens using metagenomics

Climate change, urbanization and increased interactions between animal reservoirs and human populations all increase the potential for epidemics to arise through zoonotic spillover. The vast amounts of new data generated using metagenomics[207–211], combined with AI techniques, have the potential to accelerate the discovery of new biological threats before their emergence and to predict their pandemic potential. Low-cost sequencing tools and open-source kits further support these efforts, including simplified primer designs[212].

For scenarios in which a pathogen might be shed from infected animals and samples are collected from animals or the environment, the main task after metagenomic sequencing is the identification of new pathogens for which there is no test as yet[213], followed by the prediction of the zoonotic potential of the identified pathogen sequences. While targeted sentinel surveillance methods have been used to estimate the host range of new pathogens jointly with their discovery[214,215], approaches that can be applied to routine metagenomic surveillance could help scale both pathogen discovery and spillover risk prediction. However, the utility of these approaches in identifying new pathogens and risks depends on their implementation in regions where emergence is most likely to occur.

One could envision a model for predicting the zoonotic potential of new pathogens[216], initially trained on publicly available global sequencing data. AI classification models can be pretrained on genomic sequence data and contextual information on host–pathogen interactions, which remains sparse (for example, likely cell receptor usage). Subsequently, they can be combined with environmental data (for example, urbanization and deforestation) and immunological factors (for example, likely level of existing population immunity) to generate a predictive model of pandemic potential[217–221]. For individual stakeholders performing metagenomic surveillance, the pretrained model could be hosted locally and queried to predict zoonotic potential from local metagenomic samples by including relevant environmental covariates, which could improve the model's predictive accuracy.

This scheme preserves data privacy for all stakeholders while benefiting from a shared public database. Current models for cross-species transmission define the present limits of predictive accuracy, but as with similar models trained on large amounts of data, it is possible that their accuracy might continue to improve[218,219].

Integrating multimodal data in epidemic analyses

Increasingly, epidemic analyses require the integration of diverse data types, including pathogen genomic, epidemiological, clinical, experimental, laboratory, environmental, behavioral and demographic data. For example, protein structures can now be used alongside primary genetic sequences to model evolutionary processes[93,94], while mobility data and air travel patterns are frequently used to model the risk of disease spread and the impact of interventions[5,26,95]. Given the predicted impacts of climate change, incorporating environmental variables will be useful in anticipating outbreaks and tracking spillover events[96,97].

https://doi.org/10.1038/s41591-026-04351-4

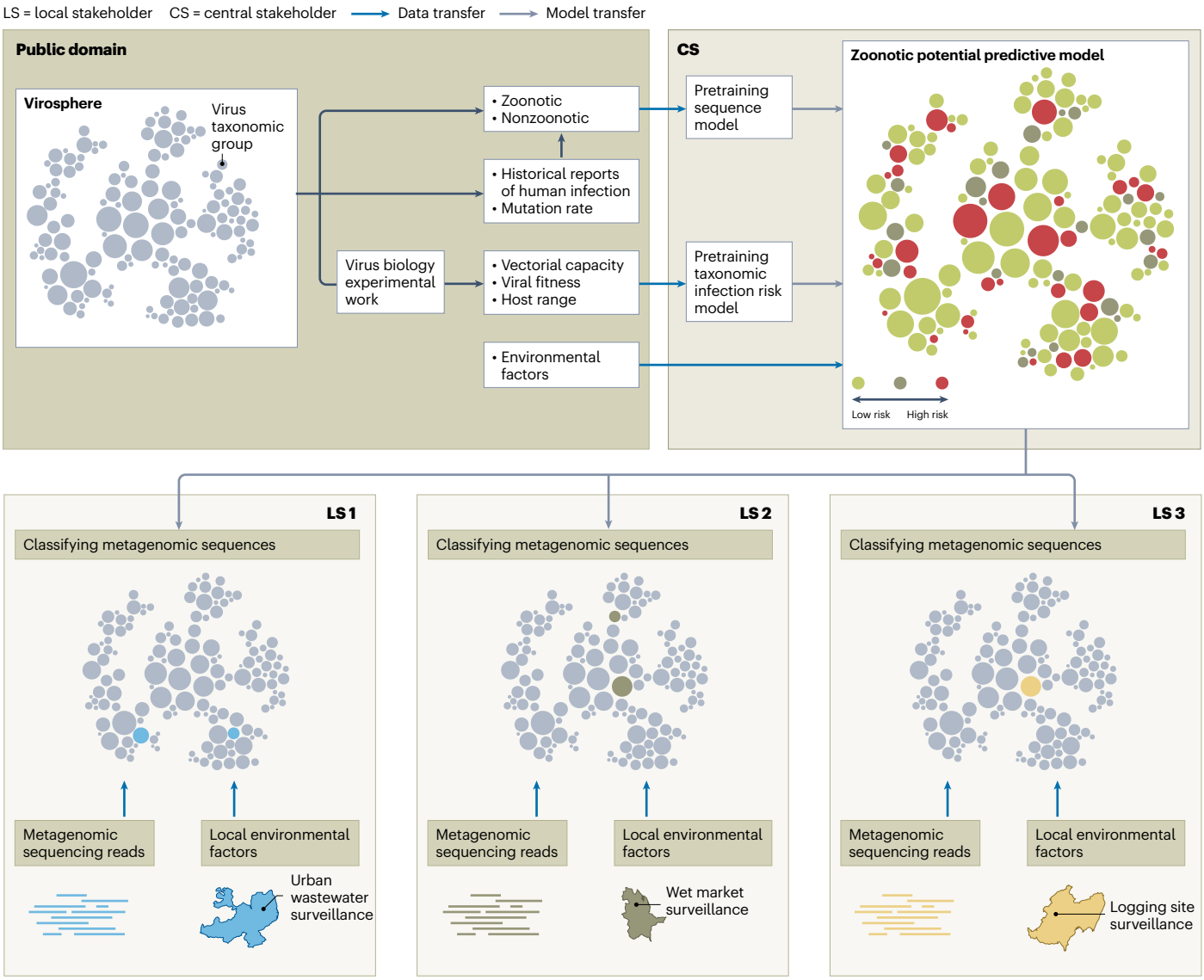

Fig. 2 | Metagenomic surveillance model for predicting zoonotic spillover of new pathogens. Current and publicly available knowledge of the virosphere, including genetic data, can be aggregated and combined with experimental work to produce information about key biological features related to cross-species transmission (top left). A central stakeholder can combine this published information to train models linking genetic sequence data (that is, genetic determinants of host specificity) and infection risk data, which are ultimately integrated into a large taxonomic zoonotic potential predictive model (top right). Local stakeholders could run local instances of the model to query locally produced metagenomic sequence data from different surveillance schemes (bottom panels).

Challenges of data integration, quality and heterogeneity

Data collected, such as mobility patterns and air traffic data, may be held by private companies and might not be publicly available during outbreaks; however, researchers and public health organizations may have existing contracts (for example, institutional sharing agreements) with such companies that allow for data use[3,26,98]. However, contracts often prohibit onward sharing with other agencies that might collect additional important data.

To overcome data-sharing limitations between the private sector, research groups and public health agencies, stakeholders might codevelop analysis pipelines using a graphical interface in which data permissions are clearly delineated. While some ad hoc analyses have already been performed, scaling them more widely requires robust software platforms and user interfaces that facilitate the integration of diverse datasets[99].

While mobility data are often commercially sensitive, climate data are more easily accessible from open-access repositories (for example, ERA5)[100], making them easier to integrate into existing frameworks. A key challenge is processing such data to match the spatial and temporal resolutions of the case and genomic data[101,102]. The DART (Dengue Advanced Readiness Tools) project provides early examples of massive spatial data integration for epidemic analyses[103], which can be easily combined with pathogen data platforms, such as the eLwazi Open Data Science Platform (https://elwazi.org/). These tools enable federated analysis of pathogen genomic data across multiple countries (South Africa, Mali, Uganda and Vietnam) with on-premises rather than cloud-based execution. For geospatial data, centralizing nonconfidential datasets on cloud-based infrastructure can facilitate coanalysis with private data, offering a pathway to ensure consistent inputs and enhance comparability across locations.

With large volumes of diverse data types, especially imagery and climate data, efficient compression algorithms are essential to reduce storage requirements and improve processing efficiency without compromising data integrity[104,105]. Methods for dimensionality reduction,

such as autoencoders (neural networks designed to learn efficient data representations), will be crucial for handling high-dimensional datasets[106–108].

The development of standardized data formats and transformation frameworks remains critical for ensuring interoperability and the ability to train models on diverse datasets[109]. Large language models (LLMs) offer opportunities for rapid mapping to existing standards, and several tools have been made available as data science capacity increases[32,37,110–112]. While we cannot expect universal standards for all data types, tools should be flexible enough to support the easy integration of federated architectures using, for example, bespoke adapters to OMOP (Observational Medical Outcomes Partnership)[43], FHIR (Fast Healthcare Interoperability Resources)[113], GA4GH (Global Alliance for Genomics and Health)[114–116] and PHA4GE (Public Health Alliance for Genomic Epidemiology)[111].

Ensuring data quality remains a major concern when attempting to aggregate and compare inferences across sites within a federated structure. This will require the creation and sharing of detailed data collection protocols, including standardized data quality control pipelines.

Infrastructure development needs

Any federated technological approach must be underpinned by the necessary policy frameworks and guiding principles, such as the FAIR (Findability, Accessibility, Interoperability and Reusability) principles[117]. These help ensure transparency, machine readability and the meaningful reuse of both data and analytical workflows. In contexts where learning is coordinated by a central stakeholder, multiple frameworks have already been developed (for example, Apache Hadoop[118], Apache Spark[119], Kubernetes[120] and Monolith[121]) to facilitate secure model training across distributed data sources, leveraging secure multiparty computation, differential privacy[122,123] and homomorphic encryption (that is, computations, analytics and data processing directly on encrypted data)[124].

Multiple techniques have been developed to reduce the risks of data breaches and reidentification. These include data minimization (which focuses on collecting only essential information), aggregation (that is, processing data as early as possible), noise injection, access controls and limiting data retention[125,126]. The *k*-anonymity[127] approach (that is, ensuring that data points cannot be distinguished from at least $k-1$ others) or synthetic data generation and digital twinning[128] (that is, creating artificial data that mimic the original in their distributions) can be used to guarantee that data are not reidentifiable. Differential privacy is a mathematical framework in which 'noise' is inserted into data and models to protect individual privacy while preserving meaningful patterns[122,123,129,130]. It has been widely adopted in distributed machine learning settings, including by major organizations such as Google[131] and the US Census[132], to protect against reconstruction attacks[133] (that is, attempts to rebuild original data from the output) or membership inference[134] (that is, determining whether a specific individual's data were used in a model).

While more efficient tools for large datasets are emerging, running these analyses locally remains impractical due to software and infrastructure limitations, highlighting the need for new methods to speed up computation[135,136]. This divide was evident during the COVID-19 pandemic, when only 42% of low- and middle-income countries reached the benchmark of sequencing 0.5% of their total reported cases, compared to 78% of high-income countries. Furthermore, low- and middle-income countries were substantially less likely to submit data within the critical 21-day window required for a real-time response[137]. Reimagining and developing shared infrastructure to support decentralized approaches, such as open-source software and cloud-based platforms for analysis and visualization, could democratize access to high-performance computing and reduce redundant development efforts. These advancements should be accompanied by skills development, expanded training opportunities and the retention of technical expertise in all settings[138–140]—a goal that will require coordinated, long-term funding[141,142]. Training and technical expertise initiatives could build upon existing and emerging partnerships. For example, the AGARI platform[143] (from Africa's Centres for Disease Control and Prevention) already harmonizes metadata and streamlines data exchange across member states, expanding sequencing and analytical capacity for outbreak responses while ensuring that countries maintain control over their own data. Leveraging these established partnerships offers an efficient pathway for training and disseminating information across diverse settings.

Integration with AI and emerging technologies

Retrieval-augmented generation, whereby LLMs access epidemiological and contextual information from data repositories, is a means of providing relevant information to commercially developed LLMs for epidemiological analyses[144]. With the increasing adoption of the Model Context Protocol (MCP)[145,146]—which enables seamless connections between AI tools and external data, tools and systems—coupled with existing approaches that leverage common data models (for example, OMOP[147] and FHIR[113]), organizations no longer need to build large infrastructures but can interact with interconnected data hubs for climate, mobility or other epidemiological data. Agentic AI workflows can then enable the rapid extraction of relevant information for answering key epidemiological questions, similar to current implementations such as AI coscientists[148]. Further, organizations can jointly contribute to training large-scale epidemiological and genomic foundation models on distributed datasets, improving the models' capabilities for understanding infectious disease transmission.

Public health equity considerations

Effective implementation requires mechanisms that guarantee the mutual benefit of federated analyses, ensuring that all participants, regardless of resources, can translate their contributions to a joint analysis into meaningful public health impact[28,149].

To promote equitable participation, decentralized analysis frameworks should align with established norms, such as the WHO's guiding principles for pathogen genome data sharing, to help ensure trusted, inclusive collaboration across jurisdictions[28]. Data and model parameter sharing, whether federated or otherwise, should enhance outbreak responses for all contributors. However, linking them directly to resource allocation (for example, across countries most affected or at risk) could improve trust[150,151]. Achieving global representativeness and coordination for such a system may require oversight by a body—such as the WHO or the European Molecular Biology Laboratory—that possesses the necessary capacity to coordinate across a diverse range of stakeholders. More broadly, developing models that minimize computational demands will increase the likelihood of complex workflows being implemented across a range of income settings.

Computational models that integrate local expertise with analytical approaches for decision-making will likely have the greatest public health impact, as observed during many past outbreaks. For example, during the COVID-19 pandemic, much of Africa's contribution to providing global early warning for variants of concern can be attributed to local capabilities, facilitated by continental capacity building and collaboration efforts[86,152,153]. More recently, in Rwanda, rapid local sequencing and analysis of the Marburg virus enabled real-time public health decision-making[154], underscoring why a parallel focus on local capacity building is paramount. Similarly, in South America, the Pan American Health Organization has established numerous initiatives, such as the Arbovirus Diagnosis Laboratory Network of the Americas (RELDA)[155], aimed at strengthening the response to arboviruses in the region. This has resulted in the timely characterization of Zika, yellow fever, dengue and chikungunya outbreaks in several countries in the region[155,156]. Extension of these initiatives toward capacity building for modeling infectious diseases is likely to have similarly important impacts parallel to the development of federated analysis frameworks.

Governance

Effective governance frameworks are essential for federated approaches to infectious disease surveillance and analysis, enabling equitable public health benefits. As federated systems distribute data stewardship and analytical responsibility across multiple actors, governance frameworks must provide clear normative guidance on roles, responsibilities, accountability and acceptable use. Such frameworks need to enable participation across diverse legal, technical and resource settings while maintaining trust, transparency and scientific integrity. Shared global public goods, including open standards, reference architectures, interoperable tools and agreed-upon practices for attribution and responsible use, are critical for lowering barriers to participation and reducing the risk of fragmentation or exclusion. Anchoring federated approaches in commonly agreed norms and accessible global infrastructure, supported by international coordination and convening mechanisms, can offer a pathway to scale federated analysis in ways that are inclusive, sustainable and responsive to global public health needs.

Conclusion

Global federated approaches to computational and genomic modeling of infectious diseases can improve outbreak detection, further our understanding of pathogen spread and help inform policy decisions to mitigate future infectious disease risks. However, although individual methods and platforms for federated analyses exist, they remain fragmented and lack a unified, accessible framework. Deploying federated analyses successfully will require unifying existing methods, developing new approaches to handle large datasets, applying robust privacy and safety precautions and oversight to analysis tools, and collaborating with stakeholders to ensure that workflows are intuitive and feasible for real-world applications. Real-time actionability is paramount; analysis frameworks must be well integrated into public health agencies and easy to use, or their utility in real-time outbreak scenarios will remain limited.

References

1. Attwood, S. W., Hill, S. C., Aanensen, D. M., Connor, T. R. & Pybus, O. G. Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. *Nat. Rev. Genet.* **23**, 547–562 (2022).
2. Mboowa, G. et al. Africa in the era of pathogen genomics: unlocking data barriers. *Cell* **187**, 5146–5150 (2024).
3. Chang, S. et al. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* **589**, 82–87 (2021).
4. Manna, A., Koltai, J. & Karsai, M. Importance of social inequalities to contact patterns, vaccine uptake, and epidemic dynamics. *Nat. Commun.* **15**, 4137 (2024).
5. Tsui, J. L.-H. et al. Genomic assessment of invasion dynamics of SARS-CoV-2 Omicron BA.1. *Science* **381**, 336–343 (2023).
6. du Plessis, L. et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708–712 (2021).
7. Khurana, M. P. et al. High-resolution epidemiological landscape from ~290,000 SARS-CoV-2 genomes from Denmark. *Nat. Commun.* **15**, 7123 (2024).
8. European Centre for Disease Prevention and Control & European Food Safety Authority. Rapid outbreak assessment—prolonged cross-border multi-serovar *Salmonella* outbreak linked to consumption of sprouted seeds. *CDC* https://www.ecdc.europa.eu/en/publications-data/rapid-outbreak-assessment-prolonged-cross-border-multi-serovar-salmonella (2025).
9. Hill, V. et al. Toward a global virus genomic surveillance network. *Cell Host Microbe* **31**, 861–873 (2023).
10. Ladner, J. T. & Sahl, J. W. Towards a post-pandemic future for global pathogen genome sequencing. *PLoS Biol.* **21**, e3002225 (2023).
11. Yozwiak, N. L., Schaffner, S. F. & Sabeti, P. C. Data sharing: make outbreak research open access. *Nature* **518**, 477–479 (2015).
12. Modjarrad, K. et al. Developing global norms for sharing data and results during public health emergencies. *PLoS Med.* **13**, e1001935 (2016).
13. Kairouz, P. et al. Advances and open problems in federated learning. *Found. Trends Mach. Learn.* **14**, 1–210 (2021).
14. Teo, Z. L. et al. Federated machine learning in healthcare: a systematic review on clinical applications and technical architecture. *Cell Rep. Med.* **5**, 101419 (2024).
15. Crowson, M. G. et al. A systematic review of federated learning applications for biomedical data. *PLOS Digit. Health* **1**, e0000033 (2022).
16. Dayan, I. et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat. Med.* **27**, 1735–1743 (2021).
17. Brisimi, T. S. et al. Federated learning of predictive models from federated Electronic Health Records. *Int. J. Med. Inform.* **112**, 59–67 (2018).
18. Sarma, K. V. et al. Federated learning improves site performance in multicenter deep learning without data sharing. *J. Am. Med. Inform. Assoc.* **28**, 1259–1264 (2021).
19. Rieke, N. et al. The future of digital health with federated learning. *NPJ Digit. Med.* **3**, 119 (2020).
20. McMahan, H. B., Moore, E., Ramage, D., Hampson, S. & Agüera y Arcas, B. Communication-efficient learning of deep networks from decentralized data. Preprint at https://doi.org/10.48550/arXiv.1602.05629 (2023).
21. Frieden, T. R., Lee, C. T., Bochner, A. F., Buissonnière, M. & McClelland, A. 7–1–7: an organising principle, target, and accountability metric to make the world safer from pandemics. *Lancet* **398**, 638–640 (2021).
22. Zwiers, L. C., Grobbee, D. E., Uijl, A. & Ong, D. S. Y. Federated learning as a smart tool for research on infectious diseases. *BMC Infect. Dis.* **24**, 1327 (2024).
23. Lyu, R., Rosenfeld, R. & Wilder, B. Federated epidemic surveillance. *PLoS Comput. Biol.* **21**, e1012907 (2025).
24. Chen, Z. et al. Global landscape of SARS-CoV-2 genomic surveillance and data sharing. *Nat. Genet.* **54**, 499–507 (2022).
25. Halabi, S., Wilder, R., Gostin, L. O. & Hurtado, M. L. Sharing pathogen genomic sequence data—toward effective pandemic prevention, preparedness, and response. *N. Engl. J. Med.* **388**, 2401–2404 (2023).
26. Tegally, H. et al. Dispersal patterns and influence of air travel during the global expansion of SARS-CoV-2 variants of concern. *Cell* **186**, 3277–3290 (2023).
27. McCrone, J. T. et al. Context-specific emergence and growth of the SARS-CoV-2 Delta variant. *Nature* **610**, 154–160 (2022).
28. *WHO Guiding Principles for Pathogen Genome Data Sharing* (World Health Organization, 2022); https://www.who.int/publications/i/item/9789240061743
29. *Attributes and Principles of Genomic Data-Sharing Platforms Supporting Surveillance of Pathogens with Epidemic and Pandemic Potential* (World Health Organization, 2025); https://www.who.int/publications/b/80650
30. Britton, T. & Scalia Tomba, G. Estimation in emerging epidemics: biases and remedies. *J. R. Soc. Interface* **16**, 20180670 (2019).
31. Li, R. et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**, 489–493 (2020).
32. Kraemer, M. U. G. et al. Tracking the 2022 monkeypox outbreak with epidemiological data in real-time. *Lancet Infect. Dis.* **22**, 941–942 (2022).
33. Kalkauskas, A. et al. Sampling bias and model choice in continuous phylogeography: getting lost on a random walk. *PLoS Comput. Biol.* **17**, e1008561 (2021).

34. Lemey, P. et al. Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. *Nat. Commun.* **11**, 5110 (2020).
35. Taylor, B. P. & Hanage, W. P. Founder effects arising from gathering dynamics systematically bias emerging pathogen surveillance. Preprint at *eLife* https://doi.org/10.7554/eLife.104201.1 (2025).
36. Vecchia, E. D. Pathoplexus: towards fair and transparent sequence sharing. *Lancet Microbe* **5**, 100995 (2024).
37. Xu, B. et al. Epidemiological data from the COVID-19 outbreak, real-time case information. *Sci. Data* **7**, 106 (2020).
38. Demonstrator video showcasing 1+MG federated analysis infrastructure—paving the way to federated learning. *European Genomic Data Infrastructure* https://gdi.onemilliongenomes.eu/news/federated-analysis-infrastructure (2025).
39. European Health Data Space Regulation (EHDS). *European Commission* https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en (2025).
40. Busch-Moreno, S. & Kraemer, M. U. G. Sequential federated analysis of early outbreak data applied to incubation period estimation. *Epidemics* **54**, 100890 (2026).
41. Deisenroth, M. & Ng, J. W. Distributed Gaussian processes. In *Proceedings of the 32nd International Conference on Machine Learning* Vol. 37 (eds Bach, F. & Blei, D.) 1481–1490 (PMLR, 2015).
42. Achituve, I., Shamsian, A., Navon, A., Chechik, G. & Fetaya, E. Personalized federated learning with Gaussian processes. In *Proc. 35th International Conference on Neural Information Processing Systems* (eds Ranzato, M. et al.) 8392–8406 (Curran Associates, 2021).
43. Hripcsak, G. et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud. Health Technol. Inform.* **216**, 574–578 (2015).
44. Schuemie, M. et al. Health-Analytics Data to Evidence Suite (HADES): open-source software for observational research. *Stud. Health Technol. Inform.* **310**, 966–970 (2024).
45. Voss, E. A. et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J. Am. Med. Inform. Assoc.* **22**, 553–564 (2015).
46. Khera, R. et al. Comparative effectiveness of second-line antihyperglycemic agents for cardiovascular outcomes: a multinational, federated analysis of LEGEND-T2DM. *J. Am. Coll. Cardiol.* **84**, 904–917 (2024).
47. Suchard, M. A. et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *Lancet* **394**, 1816–1826 (2019).
48. Schuemie, M. J., Chen, Y., Madigan, D. & Suchard, M. A. Combining Cox regressions across a heterogeneous distributed research network facing small and zero counts. *Stat. Methods Med. Res.* **31**, 438–450 (2022).
49. Xia, T., Ghosh, A., Qiu, X. & Mascolo, C. FLea: addressing data scarcity and label skew in federated learning via privacy-preserving feature augmentation. In *Proc. 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 3484–3494 (ACM, 2024).
50. Yang, Q., Liu, Y., Chen, T. & Tong, Y. Federated machine learning: concept and applications. *ACM Trans. Intell. Syst. Technol.* **10**, 1–19 (2019).
51. Abdallah, R., Abdelgaber, S. & Sayed, H. A. Leveraging AHP and transfer learning in machine learning for improved prediction of infectious disease outbreaks. *Sci. Rep.* **14**, 32163 (2024).
52. Coelho, F. C., de Holanda, N. L. & Coimbra, B. Transfer learning applied to the forecast of mosquito-borne diseases. Preprint at *medRxiv* https://doi.org/10.1101/2020.02.03.20020164 (2020).
53. Gautam, Y. Transfer learning for COVID-19 cases and deaths forecast using LSTM network. *ISA Trans.* **124**, 41–56 (2022).
54. Liu, Y., Kang, Y., Xing, C., Chen, T. & Yang, Q. A secure federated transfer learning framework. *IEEE Intell. Syst.* **35**, 70–82 (2020).
55. Roster, K., Connaughton, C. & Rodrigues, F. A. Forecasting new diseases in low-data settings using transfer learning. *Chaos Solitons Fractals* **161**, 112306 (2022).
56. Saha, S. & Ahmad, T. Federated transfer learning: concept and applications. *Intell. Artif.* **15**, 35–44 (2021).
57. Ye, Y. & Gu, A. Deep transfer learning for infectious disease case detection using electronic medical records. Preprint at https://doi.org/10.48550/arXiv.2103.06710 (2021).
58. Zhu, H., Xu, J., Liu, S. & Jin, Y. Federated learning on non-IID data: a survey. *Neurocomputing* **465**, 371–390 (2021).
59. Hsieh, K., Phanishayee, A., Mutlu, O. & Gibbons, P. B. The non-IID data quagmire of decentralized machine learning. In *Proc. 37th International Conference on Machine Learning* Vol. 119 (eds Daumé, H. & Singh, A.) 4387–4398 (PMLR, 2020).
60. Li, X., Jiang, M., Zhang, X., Kamp, M. & Dou, Q. FedBN: federated learning on non-IID features via local batch normalization. Preprint at https://doi.org/10.48550/arXiv.2102.07623 (2021).
61. Wang, J. et al. On the unreasonable effectiveness of federated averaging with heterogeneous data. Preprint at https://doi.org/10.48550/arXiv.2206.04723 (2022).
62. Reddi, S. et al. Adaptive federated optimization. Preprint at https://doi.org/10.48550/arXiv.2003.00295 (2021).
63. Karimireddy, S. P. et al. SCAFFOLD: stochastic controlled averaging for federated learning. In *Proc. 37th International Conference on Machine Learning* Vol. 119 (eds Daumé, H. & Singh, A.) 5132–5143 (PMLR, 2020).
64. Li, T. et al. Federated optimization in heterogeneous networks. In *Proc. Machine Learning and Systems* Vol. 2 (eds Dhillon, I. et al.) 429–450 (MLSys.org, 2020).
65. Collins, L., Hassani, H., Mokhtari, A. & Shakkottai, S. Exploiting shared representations for personalized federated learning. In *Proc. 38th International Conference on Machine Learning* Vol. 139 (eds Meila, M. & Zhang, T.) 2089–2099 (PMLR, 2021).
66. Arivazhagan, M. G., Aggarwal, V., Singh, A. K. & Choudhary, S. Federated learning with personalization layers. Preprint at https://doi.org/10.48550/arXiv.1912.00818 (2019).
67. Deng, Y., Kamani, M. M. & Mahdavi, M. Adaptive personalized federated learning. Preprint at https://doi.org/10.48550/arXiv.2003.13461 (2020).
68. Zhu, H. et al. FedWeight: mitigating covariate shift of federated learning on electronic health records data through patients re-weighting. *NPJ Digit. Med.* **8**, 286 (2025).
69. Li, F., Lam, H. & Prusty, S. Robust importance weighting for covariate shift. In *Proc. 23rd International Conference on Artificial Intelligence and Statistics* Vol. 108 (eds Chiappa, S. & Calandra, R.) 352–362 (PMLR, 2020).
70. Wang, V. H.-C., Lei, J., Shi, T. & Pagán, J. A. Weighting the United States All of Us Research Program data to known population estimates using raking. *Prev. Med. Rep.* **43**, 102795 (2024).
71. Yap, S. et al. Raking of data from a large Australian cohort study improves generalisability of estimates of prevalence of health and behaviour characteristics and cancer incidence. *BMC Med. Res. Methodol.* **22**, 140 (2022).
72. Chopra, A., Subramanian, J., Krishnamurthy, B. & Raskar, R. flame: a framework for learning in agent-based ModEls. In *Proc. 23rd International Conference on Autonomous Agents and Multiagent Systems* 391–399 (ACM, 2024).
73. Chopra, A. et al. Differentiable agent-based epidemiology. In *Proc. 2023 International Conference on Autonomous Agents and Multiagent Systems* 1848–1857 (ACM, 2023).

74. Quera-Bofarull, A. et al. Don't simulate twice: one-shot sensitivity analyses via automatic differentiation. In *Proc. 2023 International Conference on Autonomous Agents and Multiagent Systems* 1867–1876 (ACM, 2023).
75. Farkas, K. et al. Wastewater-based monitoring of SARS-CoV-2 at UK airports and its potential role in international public health surveillance. *PLoS Glob. Public Health* **3**, e0001346 (2023).
76. Li, J. et al. A global aircraft-based wastewater genomic surveillance network for early warning of future pandemics. *Lancet Glob. Health* **11**, e791–e795 (2023).
77. Gudde, A. et al. Predicting hospital admissions due to COVID-19 in Denmark using wastewater-based surveillance. *Sci. Total Environ.* **966**, 178674 (2025).
78. O'Reilly, K. et al. Analysis insights to support the use of wastewater and environmental surveillance data for infectious diseases and pandemic preparedness. *Epidemics* **51**, 100825 (2025).
79. Chopra, A. et al. On the limits of agency in agent-based models. In *Proc. 24th International Conference on Autonomous Agents and Multiagent Systems* 500–509 (ACM, 2025).
80. Garg, A. & Chopra, A. Distributed calibration of agent-based models. Preprint at https://openreview.net/pdf?id=tgBrJUWon5 (2024).
81. Wymant, C. et al. The epidemiological impact of the NHS COVID-19 app. *Nature* **594**, 408–412 (2021).
82. Kendall, M. et al. Drivers of epidemic dynamics in real time from daily digital COVID-19 measurements. *Science* **385**, eadm8103 (2024).
83. Baker, A. et al. Epidemic mitigation by statistical inference from contact tracing data. *Proc. Natl Acad. Sci. USA* **118**, e2106548118 (2021).
84. Chopra, A., Quera-Bofarull, A., Giray-Kuru, N., Wooldridge, M. & Raskar, R. Private agent-based modeling. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* 381–390 (International Foundation for Autonomous Agents and Multiagent Systems, 2024).
85. Chopra, A. et al. AgentTorch: large population models. *GitHub* https://github.com/AgentTorch/AgentTorch (2024).
86. Tegally, H. et al. The evolving SARS-CoV-2 epidemic in Africa: insights from rapidly expanding genomic surveillance. *Science* **378**, eabq5358 (2022).
87. Rambaut, A. et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020).
88. Hill, V. et al. A new lineage nomenclature to aid genomic surveillance of dengue virus. *PLoS Biol.* **22**, e3002834 (2024).
89. Suster, C. J. E., Pham, D., Kok, J. & Sintchenko, V. Emerging applications of artificial intelligence in pathogen genomics. *Front. Bacteriol.* **3**, 1326958 (2024).
90. Qammar, A., Karim, A., Ning, H. & Ding, J. Securing federated learning with blockchain: a systematic literature review. *Artif. Intell. Rev.* **56**, 3951–3985 (2023).
91. Jackson, C., Presanis, A., Conti, S. & De Angelis, D. Value of information: sensitivity analysis and research design in Bayesian evidence synthesis. *J. Am. Stat. Assoc.* **114**, 1436–1449 (2019).
92. Fawkes, J., Ter-Minassian, L., Ivanova, D., Shalit, U. & Holmes, C. Is merging worth it? Securely evaluating the information gain for causal dataset acquisition. Preprint at https://doi.org/10.48550/arXiv.2409.07215 (2024).
93. Malik, A. J., Poole, A. M. & Allison, J. R. Structural phylogenetics with confidence. *Mol. Biol. Evol.* **37**, 2711–2726 (2020).
94. Mifsud, J. C. O. et al. Mapping glycoprotein structure reveals *Flaviviridae* evolutionary history. *Nature* **633**, 695–703 (2024).
95. Gutierrez, B. et al. Routes of importation and spatial dynamics of SARS-CoV-2 variants during localized interventions in Chile. *PNAS Nexus* **3**, pgae483 (2024).
96. Tsui, J. L.-H. et al. Impacts of climate change-related human migration on infectious diseases. *Nat. Clim. Chang.* **14**, 793–802 (2024).
97. Carlson, C. J. et al. Pathogens and planetary change. *Nat. Rev. Biodivers.* **1**, 32–49 (2025).
98. Crawford, F. W. et al. Impact of close interpersonal contact on COVID-19 incidence: evidence from 1 year of mobile device data. *Sci. Adv.* **8**, eabi5499 (2022).
99. Brittain, J.-S. et al. GRAPEVNE—graphical analytical pipeline development environment for infectious diseases. *Wellcome Open Res.* **10**, 279 (2025).
100. Copernicus Climate Change Service, Climate Data Store. ERA5 hourly data on single levels from 1940 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)* https://doi.org/10.24381/cds.adbb2d47 (2018).
101. Moukheiber, D. et al. A multimodal framework for extraction and fusion of satellite images and public health data. *Sci. Data* **11**, 634 (2024).
102. Suel, E., Bhatt, S., Brauer, M., Flaxman, S. & Ezzati, M. Multimodal deep learning from satellite and street-level imagery for measuring income, overcrowding, and environmental deprivation in urban areas. *Remote Sens. Environ.* **257**, 112339 (2021).
103. Dasgupta, A. et al. Scalable, open-access and multidisciplinary data integration pipeline for climate-sensitive diseases. *Wellcome Open Res.* **10**, 467 (2025).
104. Kuhn, M., Kunkel, J. & Ludwig, T. Data compression for climate data. *Supercomput. Front. Innov.* **3**, 75–94 (2016).
105. Klöwer, M., Razinger, M., Dominguez, J. J., Düben, P. D. & Palmer, T. N. Compressing atmospheric data into its real information content. *Nat. Comput. Sci.* **1**, 713–724 (2021).
106. Berahmand, K., Daneshfar, F., Salehi, E. S., Li, Y. & Xu, Y. Autoencoders and their applications in machine learning: a survey. *Artif. Intell. Rev.* **57**, 28 (2024).
107. Bacchus, P., Fraisse, R., Roumy, A. & Guillemot, C. Quasi lossless satellite image compression. In *2022 IEEE International Geoscience and Remote Sensing Symposium* 1532–1535 (IEEE, 2022).
108. Liu, Y., Ponce, C., Brunton, S. L. & Kutz, J. N. Multiresolution convolutional autoencoders. *J. Comput. Phys.* **474**, 111801 (2023).
109. Zhang, C. et al. A survey on federated learning. *Knowl.-Based Syst.* **216**, 106775 (2021).
110. Gruson, H. & Jombart, T. linelist: tagging and validating epidemiological data. *Zenodo* https://doi.org/10.5281/zenodo.11954901 (2024).
111. Griffiths, E. J. et al. The PHA4GE Microbial Data-Sharing Accord: establishing baseline consensus microbial data-sharing norms to facilitate cross-sectoral collaboration. *BMJ Glob. Health* **9**, e016474 (2024).
112. Brittain, J.-S., Liggins, P. & Dasgupta, A. globaldothealth/InsightBoard. *GitHub* https://github.com/globaldothealth/InsightBoard (2024).
113. Ayaz, M., Pasha, M. F., Alzahrani, M. Y., Budiarto, R. & Stiawan, D. The Fast Health Interoperability Resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities. *JMIR Med. Inform.* **9**, e21929 (2021).
114. Rehm, H. L. et al. GA4GH: international policies and standards for data sharing across genomic research and healthcare. *Cell Genom.* **1**, 100029 (2021).
115. Thorogood, A. et al. International federation of genomic medicine databases using GA4GH standards. *Cell Genom.* **1**, 100032 (2021).
116. Fiume, M. et al. Federated discovery and sharing of genomic data using Beacons. *Nat. Biotechnol.* **37**, 220–224 (2019).

117. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
118. White, T. *Hadoop: The Definitive Guide* (O'Reilly, 2012).
119. Zaharia, M. et al. Apache Spark: a unified engine for big data processing. *Commun. ACM* **59**, 56–65 (2016).
120. Brewer, E. A. Kubernetes and the path to cloud native. In *Proc. 6th ACM Symposium on Cloud Computing* 167 (ACM, 2015).
121. Liu, Z. et al. Monolith: real time recommendation system with collisionless embedding table. Preprint at https://doi.org/10.48550/arXiv.2209.07663 (2022).
122. Dwork, C., McSherry, F., Nissim, K. & Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography* (eds Halevi, S. & Rabin, T.) 265–284 (Springer, 2006).
123. Gong, R. Exact inference with approximate computation for differentially private data via perturbations. *J. Priv. Confid.* https://doi.org/10.29012/jpc.797 (2022).
124. Rivest, R. L., Adleman, L. & Dertouzos, M. L. On data banks and privacy homomorphisms. In *Foundations of Secure Computation* 169–179 (Academic, 1978).
125. Bonawitz, K., Kairouz, P., McMahan, B. & Ramage, D. Federated learning and privacy: building privacy-preserving systems for machine learning and data science on decentralized data. *Queue* **19**, 87–114 (2021).
126. Wieland, S. C., Cassa, C. A., Mandl, K. D. & Berger, B. Revealing the spatial distribution of a disease while preserving privacy. *Proc. Natl Acad. Sci. USA* **105**, 17608–17613 (2008).
127. Sweeney, L. *k*-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **10**, 557–570 (2002).
128. Jones, D., Snider, C., Nassehi, A., Yon, J. & Hicks, B. Characterising the digital twin: a systematic literature review. *CIRP J. Manuf. Sci. Technol.* **29**, 36–52 (2020).
129. Li, T. Scalable and trustworthy learning in heterogeneous networks. *Proc. AAAI Conf. Artif. Intell.* **39**, 28715 (2025).
130. Choudhury, O. et al. Differential privacy-enabled federated learning for sensitive health data. Preprint at https://doi.org/10.48550/arXiv.1910.02578 (2020).
131. Hartmann, F. & Kairouz, P. Distributed differential privacy for federated learning. *Google Research* https://research.google/blog/distributed-differential-privacy-for-federated-learning/ (2023).
132. Abowd, J. M. The U.S. Census Bureau adopts differential privacy. In *Proc. 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 2867–2867 (ACM, 2018).
133. Dinur, I. & Nissim, K. Revealing information while preserving privacy. In *Proc. 22nd ACM SIGMOD–SIGACT–SIGART Symposium on Principles of Database Systems* 202–210 (ACM, 2003).
134. Shokri, R., Stronati, M., Song, C. & Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy)* 3–18 (IEEE, 2017).
135. Fisher, A. A. et al. Scalable Bayesian phylogenetics. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **377**, 20210242 (2022).
136. Varilly, P. et al. Delphy: scalable, near-real-time Bayesian phylogenetics for outbreaks. Preprint at *bioRxiv* https://doi.org/10.1101/2025.03.25.645253 (2025).
137. Brito, A. F. et al. Global disparities in SARS-CoV-2 genomic surveillance. *Nat. Commun.* **13**, 7003 (2022).
138. Onywera, H., Mulder, N., Kebede, Y. & Tessema, S. K. How to sustain a public-health genomics AND bioinformatics workforce in Africa. *Nat. Med.* **31**, 2480–2484 (2025).
139. Mfuh, K. O., Abanda, N. N. & Titanji, B. K. Strengthening diagnostic capacity in Africa as a key pillar of public health and pandemic preparedness. *PLOS Glob. Public Health* **3**, e0001998 (2023).
140. Shadbolt, N. et al. The challenges of data in future pandemics. *Epidemics* **40**, 100612 (2022).
141. Wong, B. L. H. et al. Harnessing the digital potential of the next generation of health professionals. *Hum. Resour. Health* **19**, 50 (2021).
142. Kaduru, C. et al. Strengthening local capacity for mathematical modelling in low- and middle-income countries: the process and lessons learnt in implementing the first cohort of Nigeria malaria modelling fellowships. *Malar. J.* **24**, 116 (2025).
143. Africa CDC launches AGARI, a continent-wide genomic data platform to strengthen outbreak response. *Africa CDC* https://africacdc.org/news-item/africa-cdc-launches-agari-a-continent-wide-genomic-data-platform-to-strengthen-outbreak-response/ (2025).
144. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* Vol. 33 (eds Larochelle, H. et al.) 9459–9474 (Curran Associates, 2020).
145. Hou, X., Zhao, Y., Wang, S. & Wang, H. Model Context Protocol (MCP): landscape, security threats, and future research directions. Preprint at https://doi.org/10.48550/arXiv.2503.23278 (2025).
146. Introducing the Model Context Protocol. *Anthropic* https://www.anthropic.com/news/model-context-protocol (2024).
147. Henke, E. et al. Conceptual design of a generic data harmonization process for OMOP common data model. *BMC Med. Inform. Decis. Mak.* **24**, 58 (2024).
148. Gottweis, J. & Natarajan, V. Accelerating scientific breakthroughs with an AI co-scientist. *Google Research* https://research.google/blog/accelerating-scientific-breakthroughs-with-an-ai-co-scientist/ (2025).
149. Karimireddy, S. P., Guo, W. & Jordan, M. I. Mechanisms that incentivize data sharing in federated learning. Preprint at https://doi.org/10.48550/arXiv.2207.04557 (2022).
150. Evertsz, N., Bull, S. & Pratt, B. What constitutes equitable data sharing in global health research? A scoping review of the literature on low-income and middle-income country stakeholders' perspectives. *BMJ Glob. Health* **8**, e010157 (2023).
151. Serwadda, D., Ndebele, P., Grabowski, M. K., Bajunirwe, F. & Wanyenze, R. K. Open data sharing and the Global South—who benefits? *Science* **359**, 642–643 (2018).
152. Viana, R. et al. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* **603**, 679–686 (2022).
153. Tegally, H. et al. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **592**, 438–443 (2021).
154. Butera, Y. et al. Genomic and transmission dynamics of the 2024 Marburg virus outbreak in Rwanda. *Nat. Med.* **31**, 422–426 (2025).
155. The Americas seek to expand genomic surveillance for dengue, chikungunya and other mosquito-borne viruses. *Pan American Health Organization* https://www.paho.org/en/news/16-8-2023-americas-seek-expand-genomic-surveillance-dengue-chikungunya-and-other-mosquito (2023).
156. Giovanetti, M. et al. Genomic and epidemiological surveillance of Zika virus in the Amazon region. *Cell Rep.* **30**, 2275–2283 (2020).
157. Madewell, Z. J., Yang, Y., Longini, I. M. Jr., Halloran, M. E. & Dean, N. E. Household secondary attack rates of SARS-CoV-2 by variant and vaccination status: an updated systematic review and meta-analysis. *JAMA Netw. Open* **5**, e229317 (2022).
158. Cuomo-Dannenburg, G. et al. Marburg virus disease outbreaks, mathematical models, and disease parameters: a systematic review. *Lancet Infect. Dis.* **24**, e307–e317 (2024).
159. Marchello, C. S. et al. Complications and mortality of non-typhoidal salmonella invasive disease: a global systematic review and meta-analysis. *Lancet Infect. Dis.* **22**, 692–705 (2022).
160. Deeks, J. J. et al. Analysing data and undertaking meta-analyses. In *Cochrane Handbook for Systematic Reviews of Interventions* (eds Higgins, J. P. T. et al.) 241–284 (Wiley, 2019).

161. Lison, A., Abbott, S., Huisman, J. & Stadler, T. Generative Bayesian modeling to nowcast the effective reproduction number from line list data with missing symptom onset dates. *PLoS Comput. Biol.* **20**, e1012021 (2024).
162. Biazzo, I., Braunstein, A., Dall'Asta, L. & Mazza, F. A Bayesian generative neural network framework for epidemic inference problems. *Sci. Rep.* **12**, 19673 (2022).
163. Semenova, E., Mishra, S., Bhatt, S., Flaxman, S. & Unwin, H. J. T. Deep learning and MCMC with aggVAE for shifting administrative boundaries: mapping malaria prevalence in Kenya. In *Epistemic Uncertainty in Artificial Intelligence* Vol. 14523 (eds Cuzzolin, F. & Sultana, M.) 13–27 (Springer, 2024).
164. Williams, R., Hosseinichimeh, N., Majumdar, A. & Ghaffarzadegan, N. Epidemic modeling with generative agents. Preprint at https://doi.org/10.48550/arXiv.2307.04986 (2023).
165. Zhang, C. & Matsen, F. A. IV. A variational approach to Bayesian phylogenetic inference. *J. Mach. Learn. Res.* **25**, 6890–6945 (2024).
166. Ki, C. & Terhorst, J. Variational phylodynamic inference using pandemic-scale data. *Mol. Biol. Evol.* **39**, msac154 (2022).
167. Chatzilena, A., van Leeuwen, E., Ratmann, O., Baguelin, M. & Demiris, N. Contemporary statistical inference for infectious disease models using Stan. *Epidemics* **29**, 100367 (2019).
168. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* **112**, 859–877 (2017).
169. Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A. & Blei, D. M. Automatic differentiation variational inference. *J. Mach. Learn. Res.* **18**, 430–474 (2017).
170. el Mekkaoui, K., Mesquita, D., Blomstedt, P. & Kaski, S. Federated stochastic gradient Langevin dynamics. In *Proc. Conference on Uncertainty in Artificial Intelligence* Vol. 161 (eds de Campos, C. & Maathuis, M.) 1703-1712 (PMLR, 2020).
171. Nemeth, C. & Fearnhead, P. Stochastic gradient Markov Chain Monte Carlo. *J. Am. Stat. Assoc.* **116**, 433–450 (2021).
172. Voznica, J. et al. Deep learning from phylogenies to uncover the epidemiological dynamics of outbreaks. *Nat. Commun.* **13**, 3896 (2022).
173. Asher, M., Lomax, N., Morrissey, K., Spooner, F. & Malleson, N. Dynamic calibration with approximate Bayesian computation for a microsimulation of disease spread. *Sci. Rep.* **13**, 8637 (2023).
174. Frazier, P. I. A tutorial on Bayesian optimization. Preprint at https://doi.org/10.48550/arXiv.1807.02811 (2018).
175. Liu, D. & Sopasakis, A. A combined neural ODE–Bayesian optimization approach to resolve dynamics and estimate parameters for a modified SIR model with immune memory. *Heliyon* **10**, e38276 (2024).
176. Reiker, T. et al. Emulator-based Bayesian optimization for efficient multi-objective calibration of an individual-based model of malaria. *Nat. Commun.* **12**, 7212 (2021).
177. He, X., Zhao, K. & Chu, X. AutoML: a survey of the state-of-the-art. *Knowl.-Based Syst.* **212**, 106622 (2021).
178. Moraga, P. et al. Bayesian spatial modelling of geostatistical data using INLA and SPDE methods: a case study predicting malaria risk in Mozambique. *Spat. Spatiotemporal Epidemiol.* **39**, 100440 (2021).
179. Lindgren, F. & Rue, H. Bayesian spatial modelling with R-INLA. *J. Stat. Softw.* https://doi.org/10.18637/jss.v063.i19 (2015).
180. Li, W., Chen, H., Jiang, X. & Harmanci, A. FedGMMAT: federated generalized linear mixed model association tests. *PLoS Comput. Biol.* **20**, e1012142 (2024).
181. Li, W. et al. Federated learning algorithms for generalized mixed-effects model (GLMM) on horizontally partitioned data from distributed sources. *BMC Med. Inform. Decis. Mak.* **22**, 269 (2022).
182. Limpoco, M. A. A., Faes, C. & Hens, N. Linear mixed modeling of federated data when only the mean, covariance, and sample size are available. *Stat. Med.* **44**, e10300 (2025).
183. Yan, Z., Zachrison, K. S., Schwamm, L. H., Estrada, J. J. & Duan, R. A privacy-preserving and computation-efficient federated algorithm for generalized linear mixed models to analyze correlated electronic health records data. *PLoS ONE* **18**, e0280192 (2023).
184. Chang, H. & Shokri, R. Bias propagation in federated learning. Preprint at https://doi.org/10.48550/arXiv.2309.02160 (2023).
185. Fowl, L., Geiping, J., Czaja, W., Goldblum, M. & Goldstein, T. Robbing the Fed: directly obtaining private data in federated learning with modified models. Preprint at https://doi.org/10.48550/arXiv.2110.13057 (2022).
186. Almodóvar, A., Parras, J. & Zazo, S. Propensity weighted federated learning for treatment effect estimation in distributed imbalanced environments. *Comput. Biol. Med.* **178**, 108779 (2024).
187. Almodóvar, A., Parras, J. & Zazo, S. Federated learning for causal inference using deep generative disentangled models. Preprint at https://openreview.net/pdf?id=r7qL5vM3Aa (2023).
188. Meurisse, M. et al. Federated causal inference based on real-world observational data sources: application to a SARS-CoV-2 vaccine effectiveness assessment. *BMC Med. Res. Methodol.* **23**, 248 (2023).
189. Vo, T. V., Lee, Y. & Leong, T.-Y. Federated causal inference from observational data. Preprint at https://doi.org/10.48550/arXiv.2308.13047 (2023).
190. Xiong, R. et al. Federated causal inference in heterogeneous observational data. *Stat. Med.* **42**, 4418–4439 (2023).
191. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D. & Shmatikov, V. How to backdoor federated learning. Preprint at https://doi.org/10.48550/arXiv.1807.00459 (2019).
192. He, M., Tang, S. & Xiao, Y. Combining the dynamic model and deep neural networks to identify the intensity of interventions during COVID-19 pandemic. *PLoS Comput. Biol.* **19**, e1011535 (2023).
193. Fu, W. et al. Privacy-preserving individual-level COVID-19 infection prediction via federated graph learning. *ACM Trans. Inf. Syst.* **42**, 82 (2024).
194. Liu, Z., Wan, G., Prakash, B. A., Lau, M. S. Y. & Jin, W. A review of graph neural networks in epidemic modeling. In *Proc. 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 6577–6587 (ACM, 2024).
195. Panja, M., Chakraborty, T., Kumar, U. & Liu, N. Epicasting: an ensemble wavelet neural network for forecasting epidemics. *Neural Netw.* **165**, 185–212 (2023).
196. Wu, Y., Yang, Y., Nishiura, H. & Saitoh, M. Deep learning for epidemiological predictions. In *41st International ACM SIGIR Conference on Research & Development in Information Retrieval* 1085–1088 (ACM, 2018).
197. Wood, D. et al. A unified theory of diversity in ensemble learning. *J. Mach. Learn. Res.* **24**, 17302–17350 (2023).
198. Mnih, V. et al. Asynchronous methods for deep reinforcement learning. Preprint at https://doi.org/10.48550/arXiv.1602.01783 (2016).
199. Samsami, M. R. & Alimadad, H. Distributed deep reinforcement learning: an overview. Preprint at https://doi.org/10.48550/arXiv.2011.11012 (2020).
200. Yin, Q. et al. Distributed deep reinforcement learning: a survey and a multi-player multi-agent learning toolbox. *Mach. Intell. Res.* **21**, 411–430 (2024).
201. Nicholls, S. M. et al. CLIMB-COVID: continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. *Genome Biol.* **22**, 196 (2021).

202. BroadE: introduction to Terra: a scalable platform for biomedical research. *Broad Institute* https://www.broadinstitute.org/videos/broade-introduction-terra-scalable-platform-biomedical-research (2021).
203. Hadfield, J. et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
204. Turakhia, Y. et al. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.* **53**, 809–816 (2021).
205. De Maio, N. et al. Maximum likelihood pandemic-scale phylogenetics. *Nat. Genet.* **55**, 746–752 (2023).
206. Kramer, A. M. et al. Online phylogenetics with matOptimize produces equivalent trees and is dramatically more efficient for large SARS-CoV-2 phylogenies than de novo and maximum-likelihood implementations. *Syst. Biol.* **72**, 1039–1051 (2023).
207. Zachariasen, T. et al. MAGinator enables accurate profiling of de novo MAGs with strain-level phylogenies. *Nat. Commun.* **15**, 5734 (2024).
208. Zhang, C., Nielsen, R. & Mirarab, S. CASTER: direct species tree inference from whole-genome alignments. *Science* **387**, eadk9688 (2025).
209. Benoit, P. et al. Seven-year performance of a clinical metagenomic next-generation sequencing test for diagnosis of central nervous system infections. *Nat. Med.* **30**, 3522–3533 (2024).
210. Wang, S. et al. PathoTracker: an online analytical metagenomic platform for *Klebsiella pneumoniae* feature identification and outbreak alerting. *Commun. Biol.* **7**, 1038 (2024).
211. Ko, K. K. K., Chng, K. R. & Nagarajan, N. Metagenomics-enabled microbial surveillance. *Nat. Microbiol.* **7**, 486–496 (2022).
212. Kent, C. et al. PrimalScheme: open-source community resources for low-cost viral genome sequencing. Preprint at *bioRxiv* https://doi.org/10.1101/2024.12.20.629611 (2024).
213. Pan American Health Organization/World Health Organization. Informative note: update cases of pneumonia due to *Legionella*—Tucumán, Argentina. *Pan American Health Organization* https://www.paho.org/sites/default/files/2023-07/20223septemberphetechnicalnotepneumonia-due-legionellaargen.pdf (2022).
214. Venkatesan, P. UK launch metagenomic pathogen surveillance programme. *Lancet Microbe* https://doi.org/10.1016/j.lanmic.2025.101143 (2025).
215. Arita, I. et al. Role of a sentinel surveillance system in the context of global surveillance of infectious diseases. *Lancet Infect. Dis.* **4**, 171–177 (2004).
216. Anker, K. M. et al. Exploring genetic signatures of zoonotic influenza A virus at the swine–human interface with phylogenetic and ancestral sequence reconstruction. *Virus Evol.* **11**, veaf028 (2025).
217. Mollentze, N., Babayan, S. A. & Streicker, D. G. Identifying and prioritizing potential human-infecting viruses from their genome sequences. *PLoS Biol.* **19**, e3001390 (2021).
218. Mollentze, N. & Streicker, D. G. Predicting zoonotic potential of viruses: where are we? *Curr. Opin. Virol.* **61**, 101346 (2023).
219. Wille, M., Geoghegan, J. L. & Holmes, E. C. How accurately can we assess zoonotic risk? *PLoS Biol.* **19**, e3001135 (2021).
220. Mollentze, N. & Streicker, D. G. Viral zoonotic risk is homogenous among taxonomic orders of mammalian and avian reservoir hosts. *Proc. Natl Acad. Sci. USA* **117**, 9423–9430 (2020).
221. Pandit, P. S. et al. Predicting the potential for zoonotic transmission and host associations for novel viruses. *Commun. Biol.* **5**, 844 (2022).

Acknowledgements

We thank all individuals who provided feedback at various stages of the project, including those at the Federated Phylodynamics and Data Integration for Early Outbreak Investigations workshop held at the KEMRI-Wellcome Trust Research Programme in Kilifi, Kenya, in April 2024. M.U.G.K. acknowledges funding from the Rockefeller Foundation (PC-2022-POP-005); Google.org; the Oxford Martin School Programmes in Pandemic Genomics & Digital Pandemic Preparedness; the European Union's Horizon Europe program projects MOOD (874850) and E4Warning (101086640); Wellcome Trust grants 303666/Z/23/Z, 226052/Z/22/Z (also to H.T., S.V.S., G.G. and J.S.B.) and 228186/Z/23/Z; UK Research and Innovation (APP8583); the Medical Research Foundation (MRF-RG-ICCH-2022-100069, also to H.T.); UK International Development (301542-403); the Bill & Melinda Gates Foundation (INV-063472); and the Novo Nordisk Foundation (NNF24OC0094346, also to H.T.). S.K.T. acknowledges a grant from the Bill & Melinda Gates Foundation (INV-018278). S. Bhatt acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/X020258/1), which is supported by the UK Medical Research Council (MRC). This UK-funded award is carried out within the framework of the Global Health EDCTP3 Joint Undertaking. S. Bhatt acknowledges support from the Danish National Research Foundation through a chair grant (DNRF160), which also supports M.P.K. and N. Scheidwasser. S. Bhatt acknowledges support from the Eric and Wendy Schmidt Fund for Strategic Innovation through the Schmidt Polymath Award (G-22-63345), which also supports M.M. S. Bhatt acknowledges support from the Novo Nordisk Foundation through the Novo Nordisk Young Investigator Award (NNF20OC0059309). S. Bhatt acknowledges support from the Novo Nordisk Foundation via the Global Pathogen Analysis Platform (GPAP) (NNF26SA0109818) which also supports H.B.H.Z. D.A.D. is supported by a Data Science-Emerging Researcher Award from the Novo Nordisk Foundation (NNF23OC0084647). This research was supported by FONIS grant SA24I0124 to L. Ferres. L. Ferres also acknowledges financial support from the Lagrange Project of the Institute for Scientific Interchange Foundation (ISI Foundation), funded by the Fondazione Cassa di Risparmio di Torino (Fondazione CRT). A.O'T., N.J.L. and A.R. acknowledge funding from ARTIC (Wellcome Trust Collaborators Award 206298/Z/17/Z ARTIC network) and ARTIC2 (Wellcome Trust Award 313694/Z/24/Z). C.H. is supported by TED's Audacious Project, including the ELMA Foundation, MacKenzie Scott, the Skoll Foundation and Open Philanthropy; National Institute of Allergy and Infectious Diseases grants U01HG007480 (H3Africa) and U54HG007480 (H3Africa); World Bank grants ACE-019 and ACE-IMPACT; the Rockefeller Foundation (grant 2021 HTH); the Africa CDC through the African Society of Laboratory Medicine (ASLM) (grant INV018978); and the Bill & Melinda Gates Foundation. N.J.L. and M.C. acknowledge support from the National Institute for Health and Care Research (NIHR) Health Protection Research Unit in Public Health Genomics. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission or other funders. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. M.A.S. is supported by US National Institutes of Health (NIH) grants AI1533044, AI162611 and AI192139.

Author contributions

M.P.K., J.L.-H.T., S.B. and M.U.G.K. conceptualized the review, with critical input from H.T. and S.V.S. B.G., M.P.K., M.A.S., E.C.H., J.L.-H.T. and M.U.G.K. created the figures. All authors critically revised the manuscript for important intellectual content. The corresponding authors affirm that all individuals listed as authors meet the appropriate authorship criteria and that no eligible contributors have been omitted.

Competing interests

M.A.S. receives contracts from Johnson & Johnson and Gilead Sciences outside the scope of this work. The other authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Samuel V. Scarpino, Samir Bhatt or Moritz U. G. Kraemer.

Peer review information *Nature Medicine* thanks Raina MacIntyre, Carl Pearson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Karen O'Leary, in collaboration with the *Nature Medicine* team.

Mark P. Khurana[1,52], Joseph L.-H. Tsui[2,3,52], Bernardo Gutierrez[2,3,4,52], Ayush Chopra[5,6], Neil Scheidwasser[1], Harrison Bo Hua Zhu[1,7,8], Serina Y. Chang[9,10,11], David A. Duchêne[1], Cathal Mills[12], Rhys P. D. Inward[2,3], Benjamin Reddy[2,12,13], John Brittain[2,14], Abhishek Dasgupta[2,14], James Sheldon[15], George Githinji[16], John S. Brownstein[17,18], Mélodie Monod[19], Luca Ferretti[2], Sivan Bershan[20,21], Simon Tietze[20], Leo Ferres[22,23], Silvia Argimón[24], Timothy J. Dallman[24], Etien Koua[25], Oliver Ratmann[26,27], Simon Cauchemez[28], Lauren A. Meyers[29], Lili Su[30], Alessandro Vespignani[23,31], Paul Pronyk[32,33], Áine O'Toole[34], Andrew Rambaut[34], Nicholas J. Loman[35], Edward C. Holmes[36], Seth Flaxman[37], Nicola Mulder[38], Oliver W. Morgan[24], Houriiyah Tegally[39], Manuel Gomez-Rodriguez[40], Nigel Shadbolt[37,41], Christian Happi[42], Meera Chand[43], Sofonias K. Tessema[44], Placide Mbala-Kingebeni[45], Marc A. Suchard[46], Oliver G. Pybus[2,3,47,53], Samuel V. Scarpino[15,48,49,50,53] ✉, Samir Bhatt[1,7,51,53] ✉ & Moritz U. G. Kraemer[2,3,53] ✉

[1]Section for Health Data Science and AI, Department of Public Health, University of Copenhagen, Copenhagen, Denmark. [2]Pandemic Sciences Institute, University of Oxford, Oxford, UK. [3]Department of Biology, University of Oxford, Oxford, UK. [4]Colegio de Ciencias Biológicas y Ambientales, Universidad San Francisco de Quito (USFQ), Quito, Ecuador. [5]MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. [6]Project Iceberg, Massachusetts Institute of Technology, Cambridge, MA, USA. [7]MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London, London, UK. [8]National Food Institute, Technical University of Denmark (DTU), Lyngby, Denmark. [9]Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA, USA. [10]UCSF UC Berkeley Joint Program in Computational Precision Health, Berkeley, CA, USA. [11]Microsoft Research, New York, NY, USA. [12]Department of Statistics, University of Oxford, Oxford, UK. [13]UKRI AI Centre for Doctoral Training in AI for the Environment, University of Oxford, Oxford, UK. [14]Doctoral Training Centre, University of Oxford, Oxford, UK. [15]Institute for Experiential AI, Northeastern University, Boston, MA, USA. [16]Kenya Medical Research Institute (KEMRI)–Wellcome Trust Research Programme (KWTRP), Kilifi, Kenya. [17]Computational Epidemiology Lab, Boston Children's Hospital, Boston, MA, USA. [18]Department of Pediatrics and Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [19]CEREMADE, CNRS, Université Paris Dauphine, Université PSL, Paris, France. [20]Exago.ml, Berlin, Germany. [21]Center for Stroke Research Berlin, Charité-Universitätsmedizin Berlin, Berlin, Germany. [22]Institute of Data Science, Faculty of Engineering, Universidad del Desarrollo, Santiago, Chile. [23]Institute for Scientific Interchange Foundation, Turin, Italy. [24]WHO Hub for Pandemic and Epidemic Intelligence, Health Emergencies Programme, World Health Organization, Berlin, Germany. [25]World Health Organization Regional Office for Africa, Brazzaville, Congo. [26]Department of Mathematics, Imperial College London, London, UK. [27]Imperial-X, Imperial College London, London, UK. [28]Mathematical Modelling of Infectious Diseases Unit, Institut Pasteur, INSERM U1332, CNRS UMR 2000, Université Paris Cité, Paris, France. [29]Department of Integrative Biology, University of Texas at Austin, Austin, TX, USA. [30]Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA. [31]Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Boston, MA, USA. [32]Centre for Outbreak Preparedness, Duke-NUS Medical School, Singapore, Singapore. [33]SingHealth Duke-NUS Global Health Institute, Duke-NUS Medical School, Singapore, Singapore. [34]Institute of Ecology and Evolution, University of Edinburgh, Edinburgh, UK. [35]Institute of Microbiology and Infection, University of Birmingham, Birmingham, UK. [36]School of Medical Sciences, The University of Sydney, Sydney, New South Wales, Australia. [37]Department of Computer Science, University of Oxford, Oxford, UK. [38]Computational Biology Division, Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town, South Africa. [39]Centre for Epidemic Response and Innovation (CERI), School for Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa. [40]Max Planck Institute for Software Systems, Kaiserslautern, Germany. [41]The Open Data Institute, London, UK. [42]Institute of Genomics and Global Health, Redeemer's University, Ede, Nigeria. [43]UK Health Security Agency, London, UK. [44]Africa Centres for Disease Control and Prevention (Africa CDC), Addis Ababa, Ethiopia. [45]Institut National de Recherche Biomédicale, Kinshasa, Democratic Republic of Congo. [46]Department of Biostatistics, University of California, Los Angeles, Los Angeles, CA, USA. [47]Department of Pathobiology and Population Sciences, Royal Veterinary College, London, UK. [48]Department of Public Health and Health Sciences, Northeastern University, Boston, MA, USA. [49]Network Science Institute, Northeastern University, Boston, MA, USA. [50]Santa Fe Institute, Santa Fe, NM, USA. [51]Pioneer Centre for Artificial Intelligence, University of Copenhagen, Copenhagen, Denmark. [52]These authors contributed equally: Mark P. Khurana, Joseph L.-H. Tsui, Bernardo Gutierrez. [53]These authors jointly supervised this work: Oliver G. Pybus, Samuel V. Scarpino, Samir Bhatt, Moritz U. G. Kraemer. ✉e-mail: s.scarpino@northeastern.edu; samir.bhatt@sund.ku.dk; moritz.kraemer@biology.ox.ac.uk